빅데이터 환경에서 개인정보 비식별화에 대한 위험성 제기 및 대응 방안 제시

이수림*, 장웅태**, 배재영***, 신찬호****, 현범수*****
*숭실대학교 경영학과
**호서전문대학교 정보보호학과
***세종대학교 정보보호학과
****고려대학교
*****인하대학교 컴퓨터정보공학과
e-mail : leeforest33@gmail.com, wkddndxo7@gmail.com,
baerow@naver.com, ssin1996@naver.com, qjatn1780@naver.com

Raising Risk and Suggesting Solution about Personal Information De-identification in Big-Data Environment

Su-Rim Lee*, Woong-Tae Jang**, Jae-Young Bae***, Chan-Ho Lee****,
Beom-Su Hyun*****
*Dept of Business Of Administration, Soong-sil University
**Dept of Information Security, Hoseo Vocational College
***Dept of Information Security, Se-jong University
****Korea University
****Dept of Computer Information Engineering, In-ha University

요 약

최근 빅데이터 산업이 발전하고 있는 상황에서 빅데이터 산업에 활용되는 개인정보의 보호에 관한 문제가 대두하고 있다. 빅데이터 산업에서 개인정보를 활용하기 위해서는 비식별화 조치를 해야 한다. 하지만 비식별화는 비식별화 평가 모델 자체의 취약성과 더불어 비식별화된 개인정보를 재식별화 하는 위험성도 존재한다. 본 논문은 적정성 평가 모델, 비식별화 조치 기술, 재식별에 관한 위험성을 연구하고 각 위험성에 대한 대응 방안을 통해 재식별화의 문제를 해결하여 빅데이터 산업에서 비식별화된 개인정보가 안전히 쓰일 수 있도록 해야 한다.

1. 서론

최근 각국의 모든 산업에서 빅데이터를 활용한 산업 발전이 이뤄지고 있다. 빅데이터를 활용하면 고객들에게 맞춤형 마케팅을 제공할 수 있다. 개인정보를 데이터베이스에 암호화하여 저장하면 데이터를 복호화 해야 하므로 효율성과 가용성이 저하되기 때문에 빅데이터를 활용하기 위해서는 개인정보를 반드시 제 3자가 알아볼 수 없도록 비식별화 조치를 해야 한다.

정부에서 2016년 6월에 발행한 『개인정보 비식별 조치 가이드라인』에서 비식별화된 개인정보를 활용하기 위해서는 적정성 평가를 거쳐야 하는데 적정성 평가 모델에 대한 위험성이 2009년부터 국내외 연구결과를 통해 제기되어 왔다. 이러한 적정성 평가 모델은 동질성 공격, 유사성 공격 등으로 인해 재식별화될 위험성이 존재한다. 그 예로 메사추세츠의 비식별화된 개인정보가 재식별된 사례가 있다.

본 논문 구성은 다음과 같다. 제2장에서는 적정성 평가 모델에 대한 개념을 소개하고. 제3장에서는 적정성 평가 모델, 비식별화 조치 기술, 재식별에 관한 위험성을 제기하고 제4장에서는 3장의 위험성에 대한 대응 방안을 제시하여 비식별화된 개인정보에 대한 안전성을 확보하고 올바른 빅데이터 활용 생태계를 조성하는 것을 목적으로 한다.

2. 적정성 평가 모델

『개인정보 비식별 조치 가이드라인』을 보면 비식별 조치된 개인정보의 안전한 수준을 판단하기 위해 적정성 평가 모델을 사용한다. 적정성 평가 모델에는 k-익명성, l-다양성 모델 등이 있다.

2.1. k-익명성 모델

k-익명성은 k 값을 정하여 k값에 따라 익명성을 보장하는 방식으로 데이터 연결공격을 예방하기 위해 2002년에 고안되었다. k-익명성은 데이터 집합에 있는 각 레코드가 적어도 k-1개의 다른 레코드와 구분되지 않도록 하여 프라이버시를 보호하는 방법이다.[1]

k값이 커질수록 익명성은 높아지나 k값이 무한대라면 결국 데이터베이스의 모든 내용의 구분이 불가능할 것이다. 결국 분석되는 데이터의 성격에 따라서 k값을 정하여 이용해야 한다.

아래 <표 1>의 공개 의료데이터를 보면 28세의 남성이 지역번호 13053인 지역에 살며 전립선염에 걸렸던 정보가 공개될 경우 <표 2>의 투표인명부를 보았을 때 아마도 그 환자는 유권자 '김민준'일 가능성이 매우 커진다.

<표 1> 공개 의료데이터

	지역코드	연령	성별	질병
1	13053	28	남	전립선염
2	13068	21	남	전립선염
3	13068	29	여	고혈압
4	13053	23	남	고혈압
5	14853	50	여	위암
6	14853	47	남	전립선염
7	14850	55	여	고혈압
8	14850	49	남	고혈압
9	13053	31	남	위암
10	13053	37	여	위암
11	13068	36	남	위암
12	13068	35	여	위암

<표 2> 선거인명부

	이름	연령	성별	지역코드
1	김민준	28	남	13053
2	박지훈	21	남	13068
3	이지민	29	여	13068
4	최현우	23	남	13053
5	정서연	50	여	14853
6	송현준	47	남	14853
7	남예은	55	여	14850
8	성민재	49	남	14850
9	윤건우	31	남	13053
10	손윤서	37	여	13053
11	민우진	36	남	13068
12	허수빈	35	여	13068

이런 데이터 연결 공격에 대응하기 위해 k-익명성 모델이 나오게 되었다. <표 3>은 각각 4개의 항목이 서로 구별되지 않기 때문에 4-익명성 처리되었다고 한다. 이렇게 될 경우 투표인명부에서 본 김민준은 어떤 그룹에 속하는지 분명하지 않게 된다. 식별 확률이 1/3로 낮아진다.

<표 3> 4-익명성 모델에 의해 익명화된 의료데이터 사례(k=4)

구분	준식별자			민감한 정보
	지역코드	연령	성별	질병
1	130**	<30	*	전립선염
2	130**	<30	*	전립선염
3	130**	<30	*	고혈압
4	130**	<30	*	고혈압
5	1485*	>40	*	위암
6	1485*	>40	*	전립선염
7	1485*	>40	*	고혈압
8	1485*	>40	*	고혈압
9	130**	3*	*	위암
10	130**	3*	*	위암
11	130**	3*	*	위암
12	130**	3*	*	위암

2.2. l-다양성 모델

l-다양성이란 k-익명성의 취약점을 보완하기 위해 나온 모델이다. k-익명성의 취약점은 '3. 위험성 제기'에서 살펴본다. l-다양성은 익명화 과정에서 충분히 다양한 l개 이상의 서로 다른 민감한 정보를 갖도록 동질 집합을 구성해야 한다.[2] 이로 인해 민감한 정보가 충분한 다양성을 가지므로 다양성의 부족으로 인한 공격에 방어할 수 있고 배경지식으로 인한 공격에도 일정 수준의 방어능력을 갖는다.

예를 들어 <표 3>에서는 k=4로 익명화 처리가 되었지만 민감한 정보인 질병은 전립선염, 고혈압으로 2종류였는데 <표 4>에서는 모든 동질 집합이 3-다양성을 통해 익명화가 되어 3개 이상의 서로 다른 민감한 정보를 갖게 되었다. <표 4>은 4-익명성 처리를 한 <표 3>에서와 같이 동일한 질병으로만 구성된 동질 집합이 존재하지 않는다.

<표 4> 3-다양성 모델에 의해 다양성 처리된 의료데이터 사례(l=3)

구분	준식별자			민감한 정보
	지역코드	연령	성별	질병
1	1305*	≦40	*	전립선염
4	1305*	≦40	*	고혈압
9	1305*	≦40	*	위암
10	1305*	≦40	*	위암
5	1485*	>40	*	위암
6	1485*	>40	*	전립선염
7	1485*	>40	*	고혈압
8	1485*	>40	*	고혈압
2	1306*	≦40	*	전립선염
3	1306*	≦40	*	고혈압
11	1306*	≦40	*	위암
12	1306*	≦40	*	위암

3. 위험성 제기

『개인정보 비식별 조치 가이드라인』에는 아래와 같은 적정성 평가 모델, 비식별화 조치 기술, 재식별에 대한 위험성이 존재한다.

3.1 적정성 평가 모델 위험성

앞에서 k-익명성 모델을 살펴보았는데 k-익명성 모델에는 여러 취약성이 존재한다.

첫 번째로 배경지식에 의한 공격은 주어진 데이터 외 공격자의 배경지식을 통해 공격 대상의 민감한 정보를 알아내는 공격이다.[3] 예를 들어 다른 데이터와의 교차로 인해 여성이 <표 3>의 레코드 1~4에 속한다는 것을 안다면 여자는 전립선염에 걸릴 수 없다는 배경지식에 따라 질병을 고혈압으로 쉽게 추정할 수 있다.

두 번째로 동질성 공격은 데이터 집합에서 동일한 민감한 정보를 이용하여 공격 대상의 민감한 정보를 알아내는 공격이다. 예를 들면 <표 3>에서 동질 집합인 레코드 5~8의 민감한 정보는 모두 '위암'이므로 k-익명성 모델이 적용되었음에도 민감한 정보가 직접적으로 노출된다.

l-다양성 평가 모델은 k-익명성 평가 모델의 취약성을 보완하기 위해 나왔음에도 취약성이 존재한다. 첫 번째로 쏠림 공격에 취약하고 두 번째로 유사성 공격에 취약하다.

쏠림 공격이란 민감한 정보가 특정한 값에 쏠려 프라이버시를 보호하지 못하는 경우를 말한다.[3] 예를 들어 <표 5>를 보면 임의의 동질 집합이 4개의 'AIDS 양성' 레코드와 1개의 'AIDS 음성' 레코드로 구성되어 있다면 민감한 정보는 2-다양성을 만족하기는 하지만 공격자는 공격 대상이 80%의 확률로 'AIDS 양성'이라는 것을 알 수 있다.

<표 5> 쏠림 공격에 취약한 사례 (k=5, l=2)

구분	준식별자		민감한 정보
	지역코드	연령	질병
1	476**	2*	AIDS 양성
2	476**	2*	AIDS 양성
3	476**	2*	AIDS 양성
4	476**	2*	AIDS 음성
5	476**	2*	AIDS 양성

유사성 공격이란 익명화된 레코드의 민감한 정보가 서로 비슷했을 때 발생하는 취약성이다. 예를 들어 동질 집합의 병명이 쏠려있지도 않고 서로 다르지만, 의미가 서로 유사할 때 <표 6>을 보면 1~3 레코드의 경우 위궤양, 급성 위염일 경우 '위'에 관련된 것이라는 것을 알 수 있다.

<표 6> 유사성 공격에 취약한 사례(k=3, l=3)

구분	준식별자		민감한 정보
	지역코드	연령	질병
1	476**	2*	위궤양
2	476**	2*	급성 위염
3	476**	2*	만성 위염
4	4790**	≧40	급성 위염
5	4790**	≧40	감기
6	4790**	≧40	기관지염

3.2 비식별화 조치 기술 위험성

기존의 비식별화 조치 기술은 원본 데이터에 있는 하나의 속성에 대해 비식별화 데이터의 하나의 비식별 속성을 만드는 '1:1' 변환방식인데 이 방법들은 한 개인과 비식별화 데이터의 한 개인이 '1:1'로 대응된다는 점 때문에 정확성은 좋지만, 재식별 될 가능성이 있다. 지금까지의 예시들의 모든 식별자에 대한 비식별화 조치는 '1:1'로 대응을 하여 재식별 될 가능성이 있는 것이다.

3.3. 재식별 위험성

서로 다른 비식별 조치를 한 데이터베이스 B와 C 사이에 겹치는 값이 존재한다면 B에 없는 정보를 C에서 얻어낼 수 있는데 이런 식으로 같은 데이터에 서로 다른 비식별 조치를 했을 때 재식별 될 가능성이 크다. 이미 해외에서는 재식별 사례가 있다.

재식별 사례로는 미국 메사추세츠 주에서 병원 방문 기록, 우편번호, 생일, 성별 등의 정보가 포함된 의료정보를 제공했다. 데이터 공개에 앞서 이름, 주소, 사회보장번호 등 식별자의 정보를 삭제하여 비식별화 조치를 하였기 때문에 개인정보가 유출되지 않을 것으로 생각하였다. 하지만 공개된 정보와 투표자 명부 정보를 조합하여 주지사에 대한 식별을 시도하였고, 그 결과 쉽게 주지사의 정보를 식별하였다.[4] 이러한 사례를 보았을 때 비식별 정보는 개인정보가 아닌 것으로 추정되지만, 새로운 결합기술이 나타나거나 결합 가능한 정보가 증가하는 경우는 정보주체가 재식별될 가능성이 있다. 따라서 비식별 정보라 하더라도 필수적인 관리적, 기술적인 보호 조치가 필요하다.

4. 대응 방안

앞서 살펴본 3가지의 위험성에 대한 대응방안을 각각 제시한다.

4.1 적정성 평가 모델 위험성에 관한 대응방안

첫 번째는 l-다양성 모델을 보완한 t-근접성 모델의 사용을 필수로 하는 것이다. 『개인정보 비식별 조치 가이드라인』에서는 비식별화에 대한 적정성 평가모델로 k-익명성만 강제로 하고 있는데 k-익명성과 l-다양성에 취약한 부분이 존재하기 때문이다.

두 번째는 완전히 새로운 익명성 모델을 제시하는 것이다. 첫 번째에 제시한 대로 t-근접성을 사용하면 안전하다고 생각할 수 있지만 전혀 그렇지 않다. t-근접성은 동질 집합에서 민감한 정보의 분포와 전체 데이터 집합에서 민감한 정보의 분포가 이하의 차이를 보여야 하는데 동질집합의 분포에 따라 전체 분포의 근삿값을 알 수 있다는 한계점이 있다.

현재 나와 있는 모델들은 서로 보완을 하여 새로운 모델로 제시되었지만 아직도 많은 한계점이 존재한다. 이러한 기존의 모델을 아울러 보완할 수 있는 새로운 익명성 모델을 제시하여 표준으로 적정성 평가를 해야 한다.

4.2 비식별 조치 위험성에 대한 대응방안

기존의 '1:1' 방식에서 'N:1' 방식의 새로운 비식별화 알고리즘을 제시한다.

모형의 종류 선택보다 모수 추정이 중요한 분야에서는 모형의 충분 통계량만을 저장하거나, 요인 분석을 통해 새로운 데이터 항목을 만드는 등의 통계적 기법을 적용한 데이터 변형·감쇄를 시도할 수 있다. 아래 <표 8>는 요인 분석의 방법 중 하나인 주성분 분석을 이용하여 식별 가능성이 있는 기존 데이터를 새로운 변수로 대체한 것을 보여준다. 주성분 분석은 변수들을 그 변수들의 통계적으로 결과에 가장 영향이 큰 선형결합(각 변수에 상수를 곱해 더한 것)으로 대체하는 것이다. 이 경우 <표 7>과 같은 데이터를 배포하려 할 때 비식별 데이터로써 <표 8>과 같은 변환된 데이터와 <표 9>에 나타나 있는 변수 변환 식을 함께 배포한다. 이렇게 생성된 변수로는 기존 값을 알아내기

어렵지만, 통계적 유용성은 비교적 잘 유지할 수 있다.

<표 7> 원본 개인정보

식별 가능 정보				민감한 정보
신장(t)	체중(w)	나이(a)	성별(s)	
156.8	51.1	25	1	독감
161.1	48.8	28	1	백혈병
181.0	72.9	31	0	간암
172.3	84.0	19	0	간암
153.1	58.2	40	1	폐암

<표 8> 요인분석으로 비식별화 조치 (N:1)

통계 처리된 변수			민감한 정보
Factor1	Factor2	Factor2	
105.6940	187.9223	51.138	독감
186.1935	359.4410	15.525	백혈병
974.5970	974.7210	753.786	간암
751.5335	427.3578	992.169	간암
810.5255	843.4920	256.041	폐암

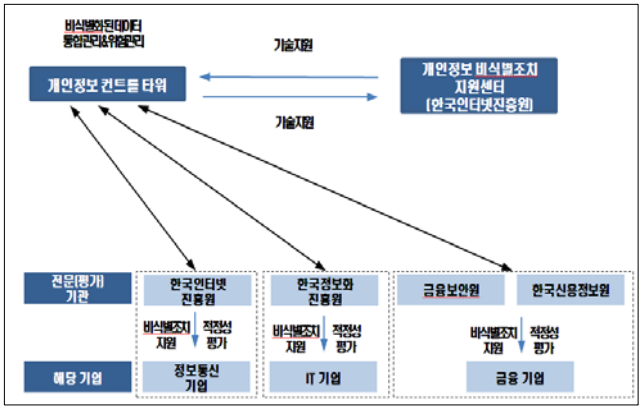
<표 9> 변수 생성 규칙

Factor1=t×6.14+w×24.58+a×36.87+s×36.87-3072.50
Factor2=t×14.38+w×8.63+a×43.16+s×8.63-3597.28
Factor3=t×5.13+w×27.54+a×1.89+s×33.21-2241.08

4.3 재식별 위험성에 대한 대응방안

시간이 지나 비식별화 된 개인정보가 데이터베이스에 많이 쌓이다 보면 결국 재식별 될 수밖에 없다. 현재 7개의 전문기관에서 비식별화 관련 모든 사항을 따로따로 관리하고 있어 어려움을 겪고 있다. 이러한 배경으로 비식별화 된 개인정보를 통합적으로 관리하고 위험 관리를 할 수 있는 개인정보 컨트롤 타워 체계를 제시한다.

(그림 1)을 보면 '비식별 조치 지원 전문기관'으로 지정된 공공기관은 한국인터넷진흥원(KISA)을 비롯해 금융보안원, 한국신용정보원, 사회보장정보원, 한국정보화진흥원이 있는데 각자 담당 분야에서 빅데이터 활용을 원하는 기업·기관의 비식별조치를 지원 및 적정성 평가를 한다. 현재 비식별화된 개인정보는 산업별로 따로 관리하여 재식별의 위험성이 커질 수 있다. 하지만 개인정보 컨트롤 타워가 존재한다면 각 전문기관으로부터 적정성 평가가 완료된 비식별화 개인정보를 받아 모든 정보를 통합적으로 관리하고 재식별의 위험성에 대해 모니터링을 하여 재식별에 대한 위험성을 각 전문기관에 통보하기 때문에 위험성을 줄일 수 있다. 또한 '개인정보 비식별조치 지원센터'는 컨트롤 타워에 기술적인 지원을 하게 되고 컨트롤 타워는 연구개발 및 모니터링 결과에 대한 자료를 제공하여 빅데이터 비식별화 관련 기술을 개선할 수 있도록 돕는다. 현실적으로 데이터가 쌓이고 시간이 지나다 보면 결국 재식별 가능성이 있을 수 있다. 이러한 개인정보 컨트롤 타워는 재식별화에 대한 가능성을 최대한으로 줄이기 위함이다.



(그림 1) 개인정보 컨트롤 타워 예시

5. 결론

지금까지 빅데이터 산업에서 개인정보 비식별화의 위험성과 재식별화 사례 그리고 대응 방안을 알아봤다. 비식별화에 대한 위험성은 다양하지만 결국에는 재식별화 문제가 가장 심각하다. 재식별화 위험성은 오래전부터 논의가 되었지만 많은 사람이 잘 인식하고 있지 않다. 사실 비식별화라는 말 자체도 재식별의 가능성을 내포하고 있는데 그렇게 때문에 재식별화에 대한 완벽한 대응 방안은 없다고 확신한다. 결국에는 빅데이터 산업에서 개인정보가 축적되다 보면 결국에는 재식별이 될 수밖에 없다.

해외의 경우 비식별화된 정보의 효율성을 고려하면서 안전성도 초점을 두는 반면 우리나라는 재식별화의 위험성은 전혀 고려하지 않고 빅데이터 활용을 위한 비식별화 방안에만 초점을 두고 있다. 이것은 정부에서 2016년 6월에 발행한『개인정보 비식별 조치 가이드라인』에서 적정성 평가 단계에서 k-익명성 모델을 사용하여 적정성을 평가하는 것만 봐도 전혀 안전성을 고려하고 있지 않다고 판단할 수 있다.

우리나라는 비식별화에 대한 위험성을 인식하여 개인정보를 빅데이터 산업에 활용하기 위한 효율성만 따지지 말고 안전성도 고려해야 한다. 위에 제시한 기술영역, 관리영역, 적정성 평가영역의 대응 방안 세 가지를 고려하면 재식별화에 대한 가능성을 최대한으로 줄일 수 있다고 확신한다. 이러한 대응 방안을 통해 재식별화의 문제를 해결하여 빅데이터 산업에서 비식별화된 개인정보가 안전히 쓰일 수 있도록 해야 한다.

참고문헌

[1] 이재식 (2013), 빅데이터 환경에서 개인정보보호를 위한 기술, Internet & Security Focus

[2] 미래창조과학부(2016), 개인정보 비식별 조치 가이드라인

[3] 한국정보화진흥원 (2014), 개인정보 비식별화에 대한 적정성 자율평가 안내서

[4] 고학수 (2015), "개인정보의 비식별화 처리가 개인정보보호에 미치는 영향에 관한 연구", 개인정보보호위원회